[D21] 沖縄語のデジタル語彙資源の構築

○宮川創 [1)]，加藤幹治 [2)]，町田星羅 [3)]，カルリノ・サルバトーレ [4)]，ズラズリ美穂 [5)]
1) 国立国語研究所，〒190-8561 東京都立川市緑町 10 番地の 2
2) 東京外国語大学・日本学術振興会
3) ハワイ大学ヒロ校
4) 九州大学・一橋大学
5) ロンドン大学東洋アフリカ学院
E-mail: so-miyagawa@ninjal.ac.jp

Making Digital Lexicon of Okinawan

MIYAGAWA So[1)], KATO Kanji[2)], MACHIDA Seira[3)], CARLINO Salvatore[4)], ZLAZLI Miho[5)]
1) National Institute for Japanese Language and Linguistics
2) Tokyo University of Foreign Studies / Japan Society for the Promotion of Science
3) University of Hawai‘i at Hilo
4) Kyushu University/Hitotsubashi University
5) SOAS University of London

【発表概要】

沖縄語は、沖縄本島で話されている日琉語族に属する北琉球諸語のうちの一言語である。国立国語研究所発行の『沖縄語辞典』（1963 年刊行、2001 年第 9 刷）は、ラテン文字を使用し、声門閉鎖音などの音素記号を一部補足したものである。沖縄語は今日に至るまで、漢字かな混じり、カタカナ、ローマ字、ひらがなのみなど様々な形で書かれてきた。本「沖縄語辞典オンライン」プロジェクトでは、まず、今までに用いられた正書法・表記法を精査し、標準的な漢字かな混じり表記法を割り出し、標準化した。次に、国語研により既に作成されている辞書のスプレッドシートデータ（XSLX）に、標準化した漢字かな混じり表記やひらがな表記、国際音声字母（IPA）を追加した上で、データをテキスト構造化の世界標準である TEI XML に変換した。さらに、この XML を変形させ、静的ウェブサイトジェネレータである Hugo を通してウェブアプリケーションを作成した。本稿ではこの「沖縄語辞典オンライン」の現在までの成果と課題について議論する。

1. はじめに

沖縄語（うちなーぐち）は、日琉語族のうちの琉球語派、北琉球語群に属する言語である。日琉語族とは、日琉祖語という祖先の言語を共有する諸言語である。日琉語族は、日本語派と琉球語派に分かれ、日本語派には、日本語共通語をはじめ、日本語諸方言と、八丈語（八丈方言）が属する。琉球語派は、北琉球語群と南琉球語群に分かれ、北琉球語群には、奄美語と沖縄語が、南琉球語群には、宮古語と八重山語と与那国語が属する。日琉諸語のそれぞれの語派、語群が分かれた年代については諸説がある。また、琉球語派を認めず、拡大東日本語派と琉球諸語と九州諸方言を含めた南日本語派に分ける考え方もある[1]。さらに、一部の学者や UNESCO は国頭語（北部沖縄、与論島、沖永良部）、徳之島語を認め、Ethnologue [2] や ISO 639-3 [3] では、与論語、沖永良部語、徳之島語、喜界島語、北部奄美語、南部奄美語を認める。言語と方言の違いに絶対的な指標がない以上、言語分類に多様性があるのは致し方ないが、本稿では、これ以上、日琉語族の分類については、述べない。本稿が主眼とするのは、沖縄語である。沖縄語は、分類によって、沖縄北部や伊江島、伊平屋島で話されている言語を入れる説と、入れない説が存在するが、本稿が主眼とする沖縄語は、琉球王国の首都であった首里の方言を中心とする。

国立国語研究所は、1963 年に『沖縄語辞典』[4]を刊行した。さらに、2001 年の『沖縄語辞典』第 9 刷が XSLX 形式でデジタルデータ化され、国立国語研究所のリポジトリに CC BY 4.0 のライセンスで公開され、現在も入手可能である[5]。

このデータセットは、沖縄語の学習や言語復興にとって大変有益であるものの、様々な補助記号を加えたアルファベットによる音韻表記が採用されており慣れが必要であること、また、データ構造に例文と意義が混在してい

ることから、学習者にとっては使いづらい。本研究は、このデータセットを活用して、沖縄語学習者にとって使いやすいウェブアプリケーションにするものである。

2. プロジェクトデザイン

2.1 『沖縄語辞典』の経緯

『沖縄語辞典』は、1947 年度文部省科学研究費共同概究題目「日本民族に近接せる諸民族の言語及び文化等の硫究」のプロジェクトにおいて、沖縄語首里方言話者である島袋盛敏氏が「琉球首里語」の題目を担当したことに端を発する。島袋氏はこの科学研究費によって、沖縄語首里方言辞典の執筆を始めた[6]。島袋氏は国立国語研究所の設立時期である 1948 年度・1949 年度に、評議員の柳田國男氏の推薦によって調査研究を外部委託され、1950 年には国立国語研究所の非常勤職員となり、その中で辞書の執筆を続けた。島袋氏の辞書の稿本は、1951 年に完成し、全体の分量は原稿用紙 1856 枚に及んだ。その後、国立国語研究所地方方言研究室が大幅に見直し、言語学的・客観的な分析をもとに、改訂され、1963 年に『沖縄語辞典』として刊行された。

島袋氏は、首里の士族階級の家庭の出身である。首里方言は特に、身分別の言語使用域の差が大きかった。しかし、明治以降の身分制度の廃止や急激な社会変動によって、差は著しく失われてしまった。島袋氏は、首里王府を支持する保守派の家庭で育ったことや、妻もまた首里方言の母語話者であること、1931 年に東京に引っ越し、それ以降、関東に住んでいたことから、首里方言の士族言葉を非常によく保っているという。

元々、島袋氏の稿本の見出し語は全てカタカナ表記でなされていたというが、その表記は、ベッテルハイム琉訳新約聖書などで使われる伝統的な仮名表記に近いものであり、実際の音韻をそのまま表したものではなかった。そこで、国立国語研究所では、研究所評議員である服部四郎教授（東大）の指導のもと、首里方言の母語話者である比嘉春潮氏らによる読み上げと加筆を基に、首里方言の音韻体系に忠実な、言語学的なアルファベット表記を用いて、見出し語および例文の表記を直した。この改訂による表記は、首里方言の音韻体系を忠実に表すという点において優れている一方、日本での日常生活で一般の人が目にすることがない補助記号が用いられているため、辞典を活用するためにはこれらの特殊な記号の発音を全て覚えなければならないという負担が存在する。

現在、しまくとぅば検定、各地の琉球諸語の教室、教科書の出版[7]など、琉球諸語の復興・保全に繋がる教育活動が盛んになっているが、沖縄語の表記法についても、ある程度現代的でかつ、古典的仮名表記よりも実際の音韻体系にある程度即した表記法が確立されつつある。沖縄語に限っては、漢字かな混じりの表記も頻繁にみられる。そのため、この『沖縄語辞典』に漢字かな表記を望む学習者の声もウェブ上で公開されている[8]。本研究では、沖縄語の表記法を可能な限り調査し、比較的標準的で、かつコンピュータでも入力しやすい、ひらがな表記と漢字かな表記を標準化して見出し語および例文を転記している。かな表記は、国語研のアルファベットから自動で変換するプログラムを作成して変換し、漢字かな表記は、意味などを参考に、3 人の作業者が漢字かな表記の入力を行なっている。

2.2 語彙資源のウェブアプリケーション化

紙の辞書では紙面の見やすさやページ数制限などを考慮する必要があるので、全ての表記法を載せることは難しい。しかし、ウェブページであれば、あらかじめ多数の表記法のデータを記録しておき、そのうち学習者が必要なものだけを使いやすいように表示させることが可能である。そして、学習者用の辞書サイトと研究者用の辞書サイトを別々に表示させることも可能である。そこで、本プロジェクトでは、2001 年の第 9 刷の『沖縄語辞典』のデータのスプレッドシートに一般的なかな表記、漢字かな混じり表記を加えた後、TEI (Text Encoding Initiative)ガイドライン P5 [9] に準拠した XML に変換し、そこから XSLT を用いてデータを変形し、ウェブサイトに落

とし込んで、言語資源として公開することを目指している。次の節では、これまでに採用してきた方法論について述べる。

3. 沖縄語辞典デジタル化の計画と進展

3.1 既存のデジタルデータと表記法

国立国語研究所はすでに『沖縄語辞典』の第 9 刷のデータをスプレッドシート（拡張子 XSLX）に入力し、それを CC BY 4.0 で公開している。『沖縄語辞典』のオリジナル版では、様々な補助記号を用いたラテン・アルファベットで見出し語と例文を表記しているが、このスプレッドシート版では、ASCII の範囲の英数字で見出し語と例文を書いているため、オリジナルではセディーユなどのダイアクリティカル・マークつきで書かれている文字が、大文字で書かれている。そこで、本プロジェクトは、『沖縄語辞典』の現物を見ながら、オリジナルのどの文字が元データのどの文字に対応するのか調査した。また、範例や説明の章においてそれらの文字の実際の発音も調べ、IPA で対応するものをできる限り調査した。そして、現在出版されている沖縄語のひらがな表記を有する書籍を調べ、ひらがな表記を標準化した。この際、パソコンでの入力しやすさを最優先した。日本語共通語にはなく沖縄語にある発音を表すために、新しいひらがなを使用する者もいるが、本プロジェクトの標準化では標準のひらがなのみを使用した。

3.2 漢字かな表記、ひらがな、IPA の追加

ひらがな表記および、IPA、そしてオリジナルの表記は、簡単な Python プログラムを用いて、『沖縄語辞典』データファイルの「見出し語」の ASCII 表記から自動変換で作成した。漢字かな混じり表記への変換は、3 名の作業者によってなされた。見出し語の漢字かな混じり表記変換は、14,549 語全てが完了し、現在チェックと推敲を行っている。しかし、例文の漢字かな混じり表記は、4 分の 1 程度しか完了していない。2022 年度中にウェブページのベータ版を公開する予定であるが、公開時点で漢字かな混じり表記を表示できるのは見出し語に限られ、例文はひらがな・IPA・オリジナル表記のみに限られると思われる。

3.3 データの TEI XML 化

『沖縄語辞典』のデータのスプレッドシートを、現在デジタルヒューマニティーズ分野においてテキストの機械可読化・構造化フォーマットの世界標準となっている TEI XML に変換し、他のプロジェクトが容易に二次利用可能な状態にする必要がある。そのために、本プロジェクトでは、Python の ElementTree ライブラリを使用して、『沖縄語辞典』のスプレッドシートデータを TEI XML に変換するプログラムを開発した。このプログラムにより変換される TEI XML ファイルは、TEI XML ガイドラインにある辞書データのサンプルおよび TEI Lex-0 [10] に従って、設計した。<orth>タグの xml:lang 属性において BCP47 に従って複数の表記を示し、各単語に固定されたピッチアクセントパターンの種類を持つ辞書を作成した。また、発音を明確にするために国際音声記号を追加し、<phon>タグを使用した。このプログラムによって変換された、辞書データの一部を図 1 で示す。

```
<entry>
  <cit>
    <bibl>『沖縄語辞典』国立国語研究所資料集5第9刷 (2009), p.100</bibl>
  </cit>
  <form>
    <orth xml:lang="ryu-Hira">あびーぐぃー</orth>
    <orth xml:lang="ryu-Jpan" n="1">叫び一声</orth>
    <orth xml:lang="ryu-Jpan" n="2">叫声</orth>
    <orth xml:lang="ryu-Latn" n="1">?abiigwii</orth>
    <orth xml:lang="ryu-Latn" n="2">abiigwii</orth>
    <pron notation="ipa">?abiigwii</pron>
    <pron notation="accent">0</pron>
  </form>
  <gramGrp>
    <pos>NOUN</pos>
    <subc>名</subc>
  </gramGrp>
  <sense xml:lang="jp-Jpan" n="1">
    <def>叫び声。kaamakara〜nu cikariiN.遠くから叫び声が聞こえる。</def>
  </sense>
  <usg></usg>
</entry>
```

図 1. 辞書データの TEI XML 化

3.4 TEI XML データのウェブサイト化

XSLT を用いて、上述のプログラムで変換された TEI XML ファイルを、ウェブサイトに変形する。この際、プロトタイプとして作成したウェブサイトには、Go 言語で開発された静的サイトジェネレータである Hugo を用いた。Hugo のテーマは、検索可能でメニューバーから文字を選んで引くことができ、それぞれの見出し語ページが見やすい、Hugo Curious を

選択した。このテーマでは、ライトモード/ダークモードをユーザが自由に選択できる。図2で、Hugo Curious を用いて開発したライトモードのサンプルウェブページを示す。

図 2. サンプルウェブページのライトモード

4. おわりに

以上、現在進行中の「沖縄語辞典オンライン」プロジェクトの進展とそこで使用している技術について詳述した。ここで述べたものは、計画され、サンプルを作成し、試しているものであり、今後の技術の革新、あるいは、プロジェクトの運営上の制限などで、技術の選定など変更される可能性がある。特に静的サイトジェネレータである Gatsby と TEI を JavaScript でウェブサイトに直接視覚化して表示させる CETEIcean を組み合わせた技術が Raffaele Viglianti 氏によって開発中であり[11]、このプログラムを用いれば、TEI を HTML に変換することなく、そのまま Gatsby のコンテンツのフォルダに入れてビルドすれば、ウェブサイトに表示される。そのため、今後は Gatsby と CETEIcean を使用することも視野に入れており、現在、本プロジェクトのサンプルデータで試している。技術は日進月歩で進化しているため、使用する技術に関しては、より適したものを使う準備がある。だが、本プロジェクトの目標は、学者向けの『沖縄語辞典』を万人がウェブ上で気軽に使えるものにして、沖縄語の言語復興に寄与することに変わりはない。今後は、沖縄語の見出し語と例文の読み上げ音声や、視覚的ハンディキャップがある人向けに、サイトの音声読み上げなど、よりインクルーシブなオンライン辞書に仕上げていく。この目標を目指し、公開に向けて作業を進めていく。

参考文献

[1] 五十嵐 陽介. 分岐学的手法に基づいた日琉諸語の系統分類の試み. フィールドと文献から見る日琉諸語の系統と歴史. 開拓社. 2021, pp. 17-51.
[2] Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). *Ethnologue: Languages of the World*. 25th ed. SIL International. 2022. http://www.ethnologue.com (参照 2022-09-26).
[3] SIL International. ISO 639-3. https://iso639-3.sil.org/code_tables/639/data (参照 2022-09-26).
[4] 国立国語研究所. 沖縄語辞典. 国立国語研究所資料集 5. 第 9 刷. 財務省印刷局. 2001.
[5] 国立国語研究所. 沖縄語辞典 データ集. https://mmsrv.ninjal.ac.jp/okinawago/ (参照 2022-09-26).
[6]以下の経緯は、『沖縄語辞典』の pp. 1–7 の「編集経過の概要」による。
[7] 例えば、花薗悟. 初級沖縄語. 研究社. 2020. や、西岡敏・仲原穣. 沖縄語の入門―たのしいウチナーグチ. 白水社. 2006. など。
[8]いめゆんな. オススメの勉強法 【うちなーぐち（沖縄方言）講座・じゅん選手】. 沖縄芸人じゅん選手のネタで！ゼロから学ぶうちなーぐち講座. 2015-07-24. https://imeyunkana.blogspot.com/2015/05/blog-post_51.html (参照 2022-09-26).
[9] TEI: Text Encoding Initiative. P5 Guidelines. https://tei-c.org/guidelines/p5/ (参照 2022-09-26).
[10]TEI Lex-0. https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html (参照 2022-09-26).
[11] Viglianti, Raffaele. raffazizzi/gatsby-ceteicean-workshop. GitHub. 2022. https://github.com/raffazizzi/gatsby-ceteicean-workshop (参照 2022-09-26).